

## Glossary

### IMG/VR Terminology

**UViG:** Uncultivated Viral Genomes are viral genomes obtained from metagenomic data. Three types of UViGs are reported from IMG/VR; (a) Contig: a single contig from a metagenome or virome identified as viral contig using the protocol according to [Paez-Espino et al 2017](#); (b) Genome Bin: a group of contigs identified as belonging to the same viral genome; in IMG/VR, genome bins are only used for giant viruses, and detected using the approach described in [Schulz et al 2020](#) (c) Integrated Provirus: UViGs identified as prophages based on [VirSorter v1.0](#), a custom detection of inoviruses as described in [Roux et al. 2020](#), or based on the detection of a host region on a viral contig by CheckV ([Nayfach et al. 2020](#)).

**vOTU:** viral Operational Taxonomic Unit, group of UViGs at the ~ species rank defined based on whole-genome average nucleotide identity. Standard thresholds of 95% ANI (average nucleotide identity) and 85% AF (aligned fraction) were used, as suggested in [Roux et al. 2019](#). Groups of two or more distinct viruses are designated as vOTU\_XXX, while singletons are designated as Sg\_XXXX.

**Quality:** The quality of UViGs is provided according to [Roux et al 2019](#). The Quality has been predicted based on completeness estimation obtained with the CheckV tool [Nayfach et al 2020](#).

**Estimated completeness:** Estimation of the UViG completeness based on similarity to known complete viral genomes and/or marker genes. This estimation reflects how much of the original genome is (estimated to be) captured in the UViG.

**VPF:** Viral Protein Family defined according to [Paez-Espino et al 2017](#)

**DTR:** Direct Terminal Repeat. Repeat regions identified on both ends of a contig, often suggesting that a complete genome was assembled. Sometimes also designated as “circular” contigs.

**NCLDV:** “NucleoCytoplasmic Large DNA Viruses”: a group of viruses also known as “giant viruses” with unusually large genomes (often 1Mb or larger). These genomes are often assembled in multiple pieces, which are then identified using genome binning approaches (see <https://doi.org/10.1038/s41586-020-1957-x>).

## IMG/VR Data sources

**EVP:** Search of public IMG metagenomes (up to 2019/12/15) using the Earth Virome Protocol ([doi.org/10.1038/nprot.2017.063](https://doi.org/10.1038/nprot.2017.063))

**CheckV:** Search of public IMG metagenomes (up to 01/04/2019) using a modified version of the Earth Virome Protocol ([doi.org/10.1101/2020.05.06.081778](https://doi.org/10.1101/2020.05.06.081778))

**Virophage:** Search of public IMG metagenomes for virophages using a custom detection pipeline (conducted in 2018, [doi.org/10.1186/s40168-019-0768-5](https://doi.org/10.1186/s40168-019-0768-5))

**Inovirus:** Search of public genomes and metagenomes for inoviruses using a custom detection pipeline ([doi.org/10.1038/s41564-019-0510-x](https://doi.org/10.1038/s41564-019-0510-x))

**VirSorter\_Curated\_Dataset:** Dataset of prophages collected in 2014 from public NCBI genomes using VirSorter v1 ([doi.org/10.7554/eLife.08490](https://doi.org/10.7554/eLife.08490))

**Giant Virus MAGs:** Search of public IMG metagenomes for nucleocytoplasmic large DNA viruses (NCLDVs), i.e. giant viruses, conducted in 2018 ([doi.org/10.1038/s41586-020-1957-x](https://doi.org/10.1038/s41586-020-1957-x))

## IMG/VR Additional annotations

In addition to the standard IMG annotation pipeline, the following tools and approaches are used to annotated IMG/VR sequences:

**Taxonomic classification:** Taxonomic assignment of IMG/VR UViGs is based on comparison of predicted cds to (i) NCBI Viral RefSeq proteins v200, and (ii) 588 marker genes identified in the VOG database v97. In addition, an affiliation to the lowest common ancestor (LCA) of vOTUs was used for any vOTU member that was not already classified based on gene content.

**Host taxonomy prediction:** Host taxonomy prediction was based on four main approaches (i) direct identification of UViGs as proviruses, (ii) sequence similarity to a microbial genome, (iii) matches to IMG CRISPR spacers, and (iv) clustering in a vOTU including UViGs with a known or predicted host.

**Predicted cds annotation:** To complement the IMG functional annotation, predicted cds from UViGs were also compared to the VPF (Viral Protein Families) database ([doi.org/10.1038/nature19094](https://doi.org/10.1038/nature19094)).

# Navigating the User Interface

IMG/VR contains viral data as Uncultivated Viral Genomes (UViGs). This data is summarized in the left most **IMG/VR Content** column of stats on the IMG/VR landing page (Figure 1) and under **Browse UViGs->Summary**. For each statistical category, the count links to a table of UViGs for that category. If the UViG count is too high to be displayed in the table, filters above the table can be used to requery the selected category using filter values. [need to add more filters...]

IMG/VR Content	
Viral Datasets	
UViG: Contig	<a href="#">1930949</a>
UViG: Genome Bin	<a href="#">2055</a>
UViG: Integrated provirus	<a href="#">100720</a>
Viral Operational Taxonomic Units (vOTUs)	
Viral Clusters	<a href="#">177336</a>
Viral Singletons	<a href="#">666236</a>
With Host	
UViG: CRISPR (near)identical	<a href="#">19229</a>
UViG: CRISPR multiple partial	<a href="#">3224</a>
UViG: Consensus vOTU host	<a href="#">50669</a>
UViG: Detection in a GVMAG	<a href="#">5</a>
UViG: Provirus	<a href="#">10033</a>
UViG: Similarity to host genome	<a href="#">104043</a>
Quality	
Genome fragment	<a href="#">1894265</a>
<50% complete	<a href="#">1751143</a>
≥50% complete	<a href="#">143122</a>
High-quality	<a href="#">116283</a>
<a href="#">Viral/Spacer BLAST</a>	

Figure 1

The data can also be explored by selecting from one of the submenus under **Browse UViGs ->**

[by Taxonomy](#)  
[by Ecosystem](#)  
[by Predicted Host](#)  
[by GeoLocation](#)  
[by Body Site](#)  
by Pfam  
by VPF %  
by Gene Count  
by Seq Length

The data can also be searched in more detail using the **Search UViGs** menu.

## Browsing the Data



Figure 2

UViGs grouped **by Taxonomy**, **by Ecosystem**, and **by Predicted Host** are displayed in a hierarchical partition viewer (Figure 2) where clicking on a category or on an item in the legend expands that category. Clicking on a breadcrumb category (displayed when hovering with the mouse over the partition) will display the list of UViGs for the hierarchy up to that category.

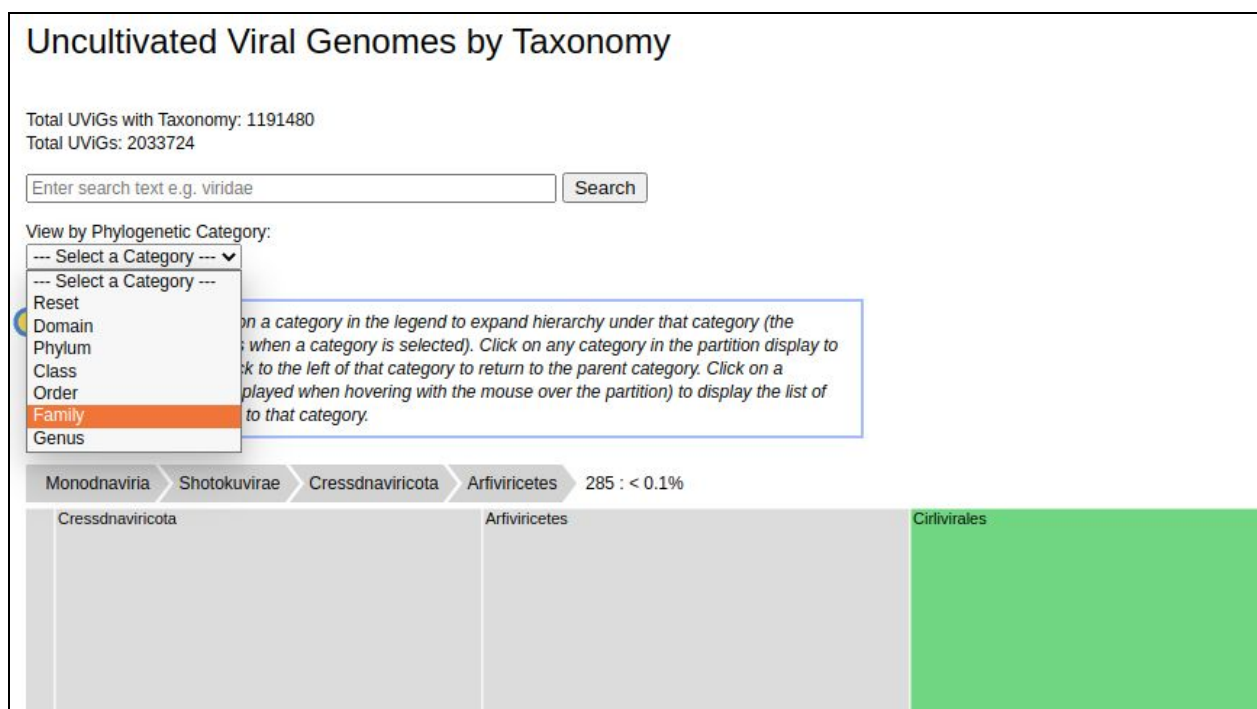


Figure 2b

Items in the partition viewer can be viewed in a table by selecting a category to view by (Figure 2b). The resultant table can be filtered by type, quality, and completeness.

## Uncultivated Viral Genomes by Phylogenetic Category

View by Phylogenetic Category:

Type:  Quality:  Completeness:

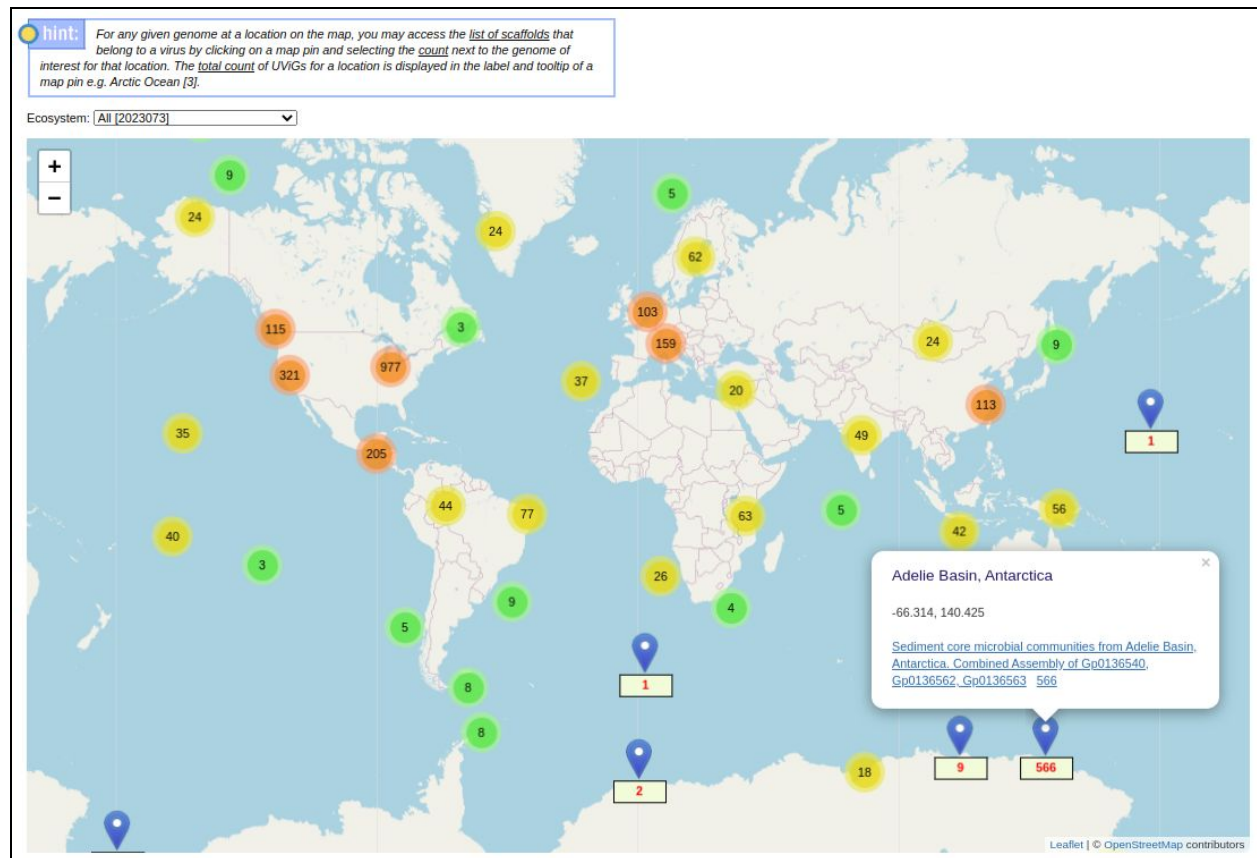
**hint:** Click on the count to view the details for selected category.  
 To group uvigs by a different category, choose from the dropdown above.  
 Select 'Reset' to go back to view all categories in the partition viewer.

Previous  2 Next      Show  entries      Search:

Family ▲	Hierarchy	UViG Count ⬆
Algavirales	Varidnaviria : Bamfordvirae : Nucleocytoviricota : Megaviricetes : Algavirales	<a href="#">21</a>
Baphyvirales	Monodnaviria : Shotokuvirae : Cressdnaviricota : Arfiviricetes : Baphyvirales	<a href="#">11</a>
Belfryvirales	Varidnaviria : Bamfordvirae : Preplasmiviricota : Tectiliviricetes : Belfryvirales	<a href="#">6</a>
Bunyavirales	Riboviria : Orthornavirae : Negarnaviricota : Ellioviricetes : Bunyavirales	<a href="#">9</a>
Caudovirales	Duplodnaviria : Heunggongvirae : Uroviricota : Caudoviricetes : Caudovirales	<a href="#">78836</a>
Chitovirales	Varidnaviria : Bamfordvirae : Nucleocytoviricota : Pokkesviricetes : Chitovirales	<a href="#">5</a>
Cirlivirales	Monodnaviria : Shotokuvirae : Cressdnaviricota : Arfiviricetes : Cirlivirales	<a href="#">267</a>
Cremevirales	Monodnaviria : Shotokuvirae : Cressdnaviricota : Arfiviricetes : Cremevirales	<a href="#">5</a>
Cryppavirales	Riboviria : Orthornavirae : Lenarviricota : Howeltoviricetes : Cryppavirales	<a href="#">14</a>
Geplafuvirales	Monodnaviria : Shotokuvirae : Cressdnaviricota : Repensiviricetes : Geplafuvirales	<a href="#">64</a>
Ghabrivirales	Riboviria : Orthornavirae : Duplornaviricota : Chrymotiviricetes : Ghabrivirales	<a href="#">2</a>

Figure 2c

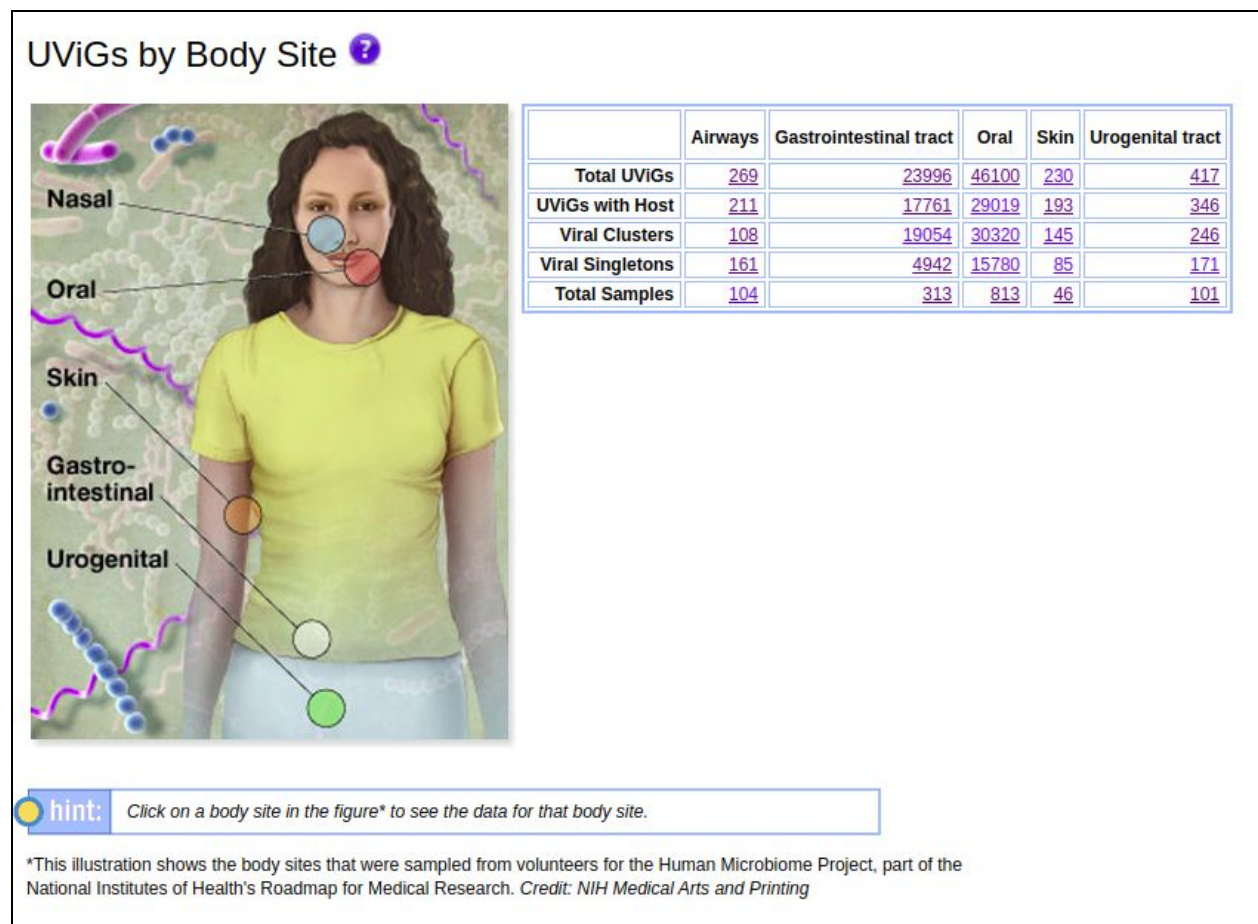
UViGs with ecosystem information and location coordinates can be viewed on a map and analyzed based on their geographic location. Browse UViGs -> **by GeoLocation** (Figure 3).



**Figure 3**



UViGs associated with a body site, can also be viewed grouped **by Body Site** (Figure 4).



**Figure 4**

UViGs grouped **by VPF %**, **by Gene Count**, and **by Seq Length** are displayed as bar charts with categories of various ranges.

## Searching the Data

UViG data can be queried using a comma-separated list of IDs, using a [combination of UViG attributes](#) (Figure 5), using a comma-separated list of scaffold IDs, or by using BLAST.

## Uncultivated Viral Genome (UViG) Search

[Search Uncultivated Viral Genome \(UViG\) by ID](#)

<b>Length Range (bases)</b>	From <input style="width: 100%;" type="text"/>	To <input style="width: 100%;" type="text"/>
<b>Gene Count</b>	From <input style="width: 100%;" type="text"/>	To <input style="width: 100%;" type="text"/>
<b>Completeness (0 - 100)</b>	From <input style="width: 100%;" type="text"/>	To <input style="width: 100%;" type="text"/>
<b>vOTU (list)</b>	<input style="width: 100%;" type="text"/> example: Sg_097066, vOTU_000462, etc.	
<b>Ecosystem</b>	<input style="width: 100%;" type="text"/> Type search term, then hit <Enter> <input type="checkbox"/> Engineered <input type="checkbox"/> Environmental <input type="checkbox"/> Host-associated	
<b>Host</b>	<input style="width: 100%;" type="text"/> Type search term, then hit <Enter> <input type="checkbox"/> Archaea <input type="checkbox"/> Bacteria <input type="checkbox"/> Bamfordvirae	
<b>Taxonomy</b>	<input style="width: 100%;" type="text"/> Type search term, then hit <Enter> <input type="checkbox"/> Unknown <input type="checkbox"/> Duplodnaviria <input type="checkbox"/> Monodnaviria <input type="checkbox"/> Riboviria <input type="checkbox"/> Varidnaviria	
<b>Body Site</b>	<input style="width: 100%;" type="text"/>	
<b>Percent VPF (0 - 100)</b>	From <input style="width: 100%;" type="text"/>	To <input style="width: 100%;" type="text"/>
<b>Pfam ID (list)</b>	<input style="width: 100%;" type="text"/> example: pfam02801, pfam00107, etc.	

**Figure 5**

## UViG Details

The details page for a UViG provides all the known information about the UViG (Figure 6), displays the neighborhood of the genes belonging to this UViG on its scaffold source (Figure 7), the predicted host information (Figure 8), taxonomy information (Figure 9), and the genes for this UViG listed in a table (Figure 10). In addition, the user can [search for similar UViGs](#) using a selected cut-off value to view a plot (Figures 11, 12).



### Uncultivated Viral Genome (UViG) Detail

UViG: IMGVR\_UViG\_2013515002\_000001

UViG Detail

Neighborhood

Host Prediction

Taxonomy

Genes in UViG

Similar UViGs

Uncultivated Viral Genome Information	
UViG ID	IMGVR_UViG_2013515002_000001
UViG Type	Integrated provirus
Percent VPF	27.27
Estimated Completeness (%)	21.59
Quality	Genome fragment
Viral OTUs	<a href="#">vOTU_807124</a>
Total Length	6476
Total Gene Count	11 <a href="#">see complete list</a>
tRNA Gene Count	0
CDS Gene Count	11
UViG Origin	<a href="#">Inovirus</a>
Source	
Metagenome	<a href="#">Hot spring microbial communities from Yellowstone National Park, Wyoming, USA - YNP13</a> <a href="#">Bechler Spring</a>
Scaffold(s)	<a href="#">YNP13_C5032</a> 2945..9420
Ecosystem	
Ecosystem	Environmental
Ecosystem Category	Aquatic
Ecosystem Type	Thermal springs
Ecosystem Subtype	Hot (42-90C)
Habitat	Thermal_springs
Body Site	

### Figure 6

### Uncultivated Viral Genome (UViG) Detail

UViG: IMGVR\_UViG\_2013515002\_000001

UViG Detail

Neighborhood

Host Prediction

Taxonomy

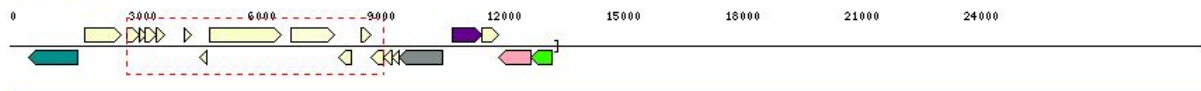
Genes in UViG

Similar UViGs

## UViG Neighborhood

Color UViG By: ☐ COG ☐ GC ☐ KEGG ☒ Pfam ☐ TIGRfam ☐ Phylo Distribution

[YNP13\\_C5032](#) [Show Color Scheme](#)



### Figure 7

## Uncultivated Viral Genome (UViG) Detail

UViG: IMGVR\_UViG\_2013515002\_000001

UViG Detail Neighborhood Host Prediction Taxonomy Genes in UViG Similar UViGs

### Host Prediction

Showing 1 to 3 of 3 entries

First Previous 1 Next Last Export Show 10 Search:

Host Detection Method	Spacer ID	Spacer Sequence	Percent Identity	Host Taxon	Host Scaffold	Host Domain	Host Phylum	Host Class	Host Order	Host Family	Host Genus	Host Species
Consensus vOTU host						Bacteria	Aquificae	Aquificae	Aquificales	Aquificaceae	Thermocrinis	
Provirus				<a href="#">Thermocrinis Tcrinis-r02</a>	<a href="#">Tcrinisr02_Tcrinis-r02_scf1119030634475.123</a>	Bacteria	Aquificae	Aquificae	Aquificales	Aquificaceae	Thermocrinis	
Similarity to host genome						Bacteria	Aquificae	Aquificae	Aquificales	Aquificaceae	Thermocrinis	

First Previous 1 Next Last Export

Showing 1 to 3 of 3 entries

Figure 8

## Uncultivated Viral Genome (UViG) Detail

UViG: IMGVR\_UViG\_2013515002\_000001

UViG Detail Neighborhood Host Prediction Taxonomy Genes in UViG Similar UViGs

### UViG Lineage

Showing 1 to 1 of 1 entries

First Previous 1 Next Last Export Show 10 Search:

Method	Domain	Phylum	Class	Order	Family	Subfamily	Genus	Species
VOGdb markers	Monodnaviria	Loebvirae	Hofneiviricota	Faserviricetes	Tubulavirales	Inoviridae		

First Previous 1 Next Last Export

Showing 1 to 1 of 1 entries

Figure 9

## Uncultivated Viral Genome (UViG) Detail

UViG: IMGVR\_UViG\_2013515002\_000001

UViG Detail Neighborhood Host Prediction Taxonomy Genes in UViG Similar UViGs

### Genes in UViG

Scaffold: [YNP13\\_C5032](#)

Add Selected to Gene Cart

Select All

Clear All

Filter column: Start Coordinate Filter text: Apply ?

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Gene ID	Locus Type	Gene Name	Start Coordinate	End Coordinate	Strand	Pfams
<input type="checkbox"/>	<a href="#">YNP13_155140</a>	CDS	MiaB-like tRNA modifying enzyme	482	1708	(-)	<a href="#">pfam00919</a> <a href="#">pfam04055</a>
<input type="checkbox"/>	<a href="#">YNP13_155150</a>	CDS	hypothetical protein	1871	2794	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155160</a>	CDS	hypothetical protein	2945	3262	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155170</a>	CDS	hypothetical protein	3262	3390	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155180</a>	CDS	hypothetical protein	3401	3694	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155190</a>	CDS	hypothetical protein	3691	3897	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155200</a>	CDS	hypothetical protein	4396	4539	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155210</a>	CDS	hypothetical protein	4775	4912	(-)	
<input type="checkbox"/>	<a href="#">YNP13_155220</a>	CDS	hypothetical protein	5029	6843	(+)	
<input type="checkbox"/>	<a href="#">YNP13_155230</a>	CDS	Zonula occludens toxin	7071	8159	(+)	<a href="#">pfam05707</a>
<input type="checkbox"/>	<a href="#">YNP13_155240</a>	CDS	hypothetical protein	8269	8604	(-)	

Figure 10

## Uncultivated Viral Genome (UViG) Detail

UViG: IMGVR\_UViG\_2013515002\_000001

UViG Detail Neighborhood Host Prediction Taxonomy Genes in UViG Similar UViGs

### Find Similar Viruses and UViGs

Find similar viruses and UViGs based on Pfams.

Cut-off Value: 0.3

Find

Plot

[Data will open in a new Window or Tab]

Figure 11

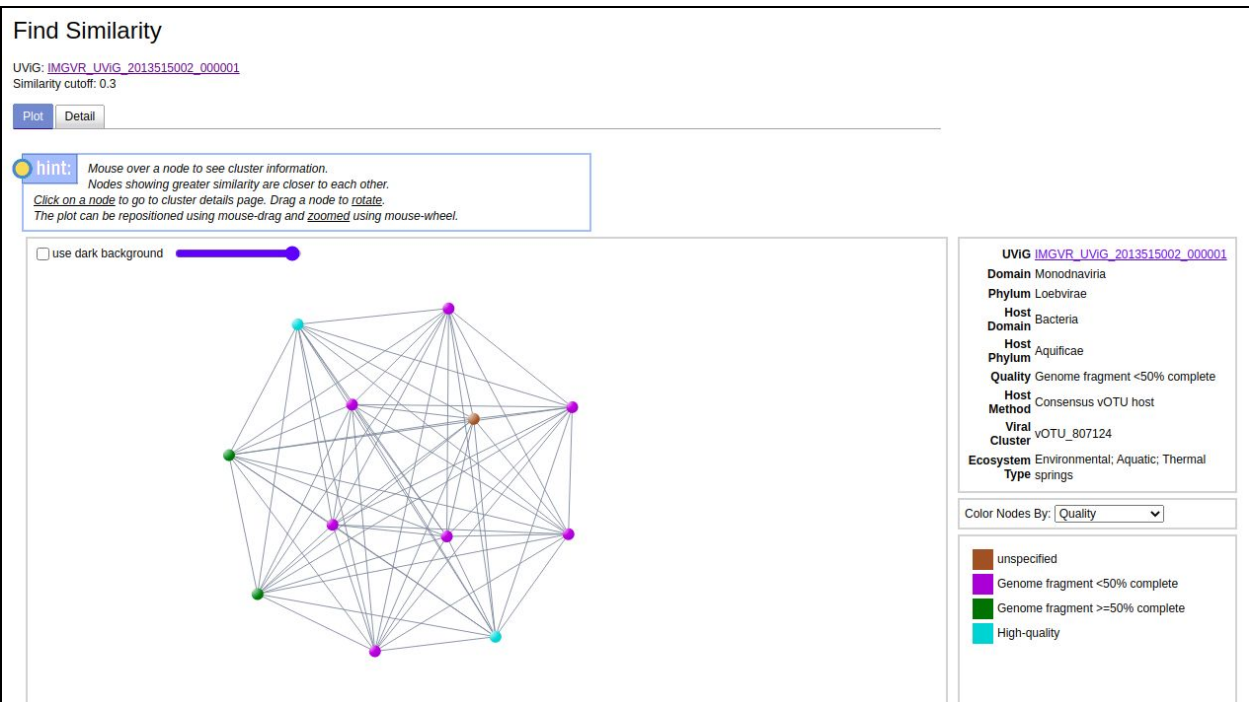


Figure 12

